A Chinese Sign Language Recognition Researh Based on Depth Image Information

Yang Quan

Software Institute, Xi'an University of Arts and Science
Xi'an 710127, China

Yangquan1110@yeah.net

Abstract

An improved Depth Image CamShift (DI_CamShift) algorithm and SIFT-BoW feature are proposed to realize the accurate tracking of gestures in the sign language video and effective matching. First, it used Kinect to obtain the depth image information of sign language gestures. Second, it adjusted the search window by calculated spindle direction angle and mass center position of the depth images. Then the calculation of minimum depth information value in the search window was used to determine the target gesture area. Finally, SIFT-BoW is calculated for gesture matching and recognition. Experiments results show that the algorithm has good robustness and can effectively track and match the sign language gestures.

Keywords

DI_CamShift; Kinect; Depth Image; SIFT-BoW

Introduction

Sign language is a natual way of information exchange and communication used by the hearing impaired with the hand shape, movement of wrist and arm, corresponding facial expressions and lip shape of pronunciation, as well as other body postures. Chinese sign language is composed of finger spelling and hand gesture. Finger spelling uses 30 letters of manual alphabet as its basic unit and spell out words in accordance with the order and rules of Chinese pinyin while each of the letters is presented through finger changing and action. Depends on indicative finger gestures, hand gesture simulates shape of object and movement to express its meaning well. But it is difficult to present Chinese characters' meaning comprehensively and accurately by hand gestures because it includes a lot of words. Compared with it, finger spelling can express many professional terms and abstract concept owing to imitate the spelling way of pinyin which bring about the characteristics of easy learning and lesser gestures. Therefore, manual alphabet recognition is an important part of Chinese sign language recognition. The current Chinese sign language implementation plan is released by the Ministry of Education and the Chinese Characters Reform Commission, which includes 30 letters of manual alphabet. As shown in Figure 1, it has 26 single letters from *A* to *Z*, 4 double letters include *ZH*, *CH*, *SH*, *NG*.

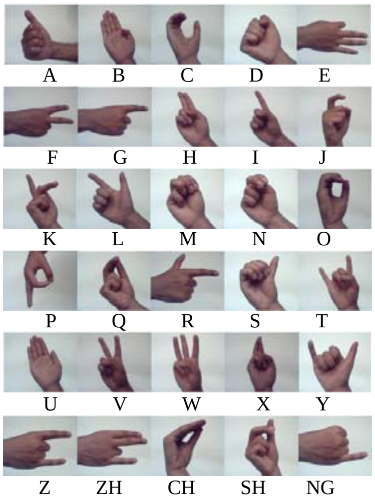


FIG. 1 CHINESE SIGN LANGUAGE ALPHABET

Sign language recognition can be divided into two categories: (a) sign language recognition based on machine vision. (b) sign language recognition based on wearable devices. Generally, wearable devices include location tracker and data gloves can provides accurate data of hand moving, but this method is difficult for application because of its high cost and complicated way of using. With the popularization and application of video acquisition device, more and more researchs focus on the computer vision based methods depend on its advantages of natural and convenient way of interaction, especially on the aspect of low cost of equipment makes it very suitable for practical application.

Kinect, short for the Kinect for Xbox360, is a motion sensing input device by Microsoft for the Xbox 360 video game console and Windows PCs. Based around a webcam-style add-on peripheral for the

Xbox 360 console, it enables users to control and interact with the Xbox 360 without the need to touch a game controller. The Kinect sensor is a horizontal bar connected to a small base with a motorized pivot and is designed to be positioned lengthwise above or below the video display. The device features an "RGB camera, depth sensor and multi-array microphone running proprietary software". The depth sensor consists of an infrared laser projector combined with a monochrome CMOS sensor, which captures video data in 3D under any ambient light conditions. So Kinect is a 3D multifunction camera and can get color images and 3D depth information at the same time.

Using Kinect as video acquisition device in the machine vision based sign language research, it can get the depth image information of gesture corresponding with color gesture video. This paper improved classic CamShift algorithm by use the Kinect depth image information, and extracted the SIFT-BoW feature of gestures for matching and recognition of Chinese sign language letters.

DI_CamShift Algorithm Based on Depth Image Information

When tracking the hand, the global operation of CamShift algorithm in the process of reverse projection increases unnecessary computation burden, thus may reduce the tracking performance. The main reason is that brightness of background region is not strong, so the reverse projection will produce noise. When converse from RGB space to HSV space, the pixel with lower brightness and saturation shows the low stability, and make some irrelevant points to be the target area. It's also possible to include irrelevant objects in the tracking window while tracking model is based on skin color. Some further research based on single hand only detect and tracking the right hand as the target object· but when other moving object· appearing in the video under the complex background, CamShift algorithm may have false tracking problem, and the misjudgment can be very obvious especially in the situation of other hand moving simultaneously.

Because the depth image took by Kinect contains information related to distance data between Kinect and its scene object's surface, therefore it is more intuitive to show 3D features of the object's surface without Interferences of colors, brightness of shadow problems compared with color image. According to the definition of depth image, it was characterized as follows: (1) Color-blind, depth image is different from color image in which it barely be affected by light, shadow and changes of surrounding environment. (2) The change' s direction of the depth image grey value is in accordance with the Z direction of view, that means 3D space can be reconstructed within the feasible region by the depth image. It can also be used to solve the problem of shade or overlap on the basis of properties belonging to the depth image. If two objects are blocked, the stratified condition of grey value generated by their different distance from the camera can be used to distinguish them. As long as set a threshold value to separate two objects with the relationship of front and behind, occlusion problem can be resolved contrast with optical image.

In consideration of insufficient transformation and tracking results given by the CamShift algorithm in color space, this paper uses an improved CamShift algorithm based on depth image information, which named Depth Image CamShift (DI_CamShift) algorithm.

D(x, y) is the depth image, and its (p + q) order 2D origin moment is defined as:

$$M_{PQ} = \sum_x \sum_y x^P y^q D(x, y) \qquad p, q = 0,1,2,\ldots\ldots \quad (1)$$

where D(x, y) indicates the depth value of pixel in the location of (x, y).

μ_{pq} is the (p + q) order central moment of D(x, y). It's defined as:

$$\mu_{pq} = \sum_x \sum_y (x-\bar{x})^p (y-\bar{y})^q D(x, y) \quad (2)$$

Its second central moment can be used as the spindle of sign language gesture in the frame, and the spindle direction is determined by the directions of the maximum and minimum second moment, that are major axis and minor axis. According to the theory of moment, the angle of spindle direction θ is calculated by:

$$\theta = \frac{1}{2}\tan^{-1}\frac{2\mu_{11}}{\mu_{20}-\mu_{02}} \quad (3)$$

In formula (3), θ is the angle between spindle and coordinate axis, its scope ranged in $\left[-\frac{\pi}{4}, \frac{\pi}{4}\right]$ which shown in the table below.

TABLE I ANGLE BETWEEN SPINDLE OF SIGN GESTURE AND AXIS

μ11	μ20-μ02	θ
0	-	0
+	-	-π/ 4<θ<0
+	0	0
+	+	0<θ<π/ 4
0	0	0
-	+	-π/ 4<θ<0
-	0	0
-	-	0<θ<π/ 4

If θ is the spindle direction of sign language gesture S, then

$$S^2(\theta)=\frac{1}{n}[(S_1(\theta)-m)^2+(S_2(\theta)-m)^2+\ldots+(S_n(\theta)-m)^2] \qquad (4)$$

where $m=\frac{1}{n}[S_1(\theta)+S_2(\theta)+\ldots+S_n(\theta)]$, $S_i(1,2,3,\ldots,n)$ is the spindle direction of object extract from the average frames with same sign language gesture.

Specific steps for DI_CamShift are:

(1) Set the entire depth image as search area.

(2) Use frame difference method to detect the area of moving hand and initialize the search window, locate its size and position.

(3) Calculate the probability distribution of depth histogram in the Search Window area.

(4) Calculate θ_1 and θ_2 separately, they are the major axis and minor axis directions of gesture in the depth image.

(5) Apply MeanShift algorithm to calculate the mass center of depth gesture image in the window, adjust the size of Search Window according to position of the mass centre and the spindle direction θ_1 and θ_2.

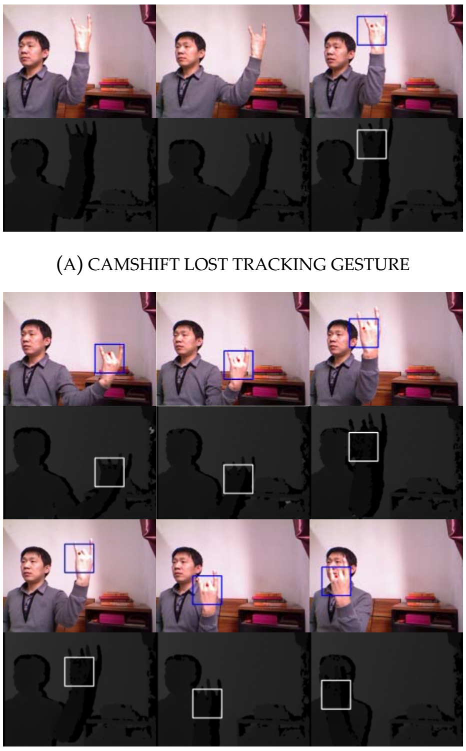
(6) For the next sign language video frame, use the mass centre and size of the Search Window generated by (3) and jump to (3) to continue.

(7) If multiple moving targets are detected, the real sign gesture is HandGesture=Min{$M_{00}(Obj_1)$, $M_{00}(Obj_2)$, … , $M_{00}(Obj_n)$}. The closer to camera, the greater depth value can be obtained by object. As the sign language gesture in front of a signer's body is considered to be the target hand and the closest object to Kinect camera, Search Window with minimum value of zero order moment contains the minimum sum of depth value, which comes from the pixels of gesture. Thus the window can be identified as the top target area.

Once confirmed the tracking window in depth video, it will be drawn to the corresponding color video in same location simultaneously for tracking.

Under the same scene, DI_CamShift algorithm has better performance of tracking, not only avoid lost tracking but also correct the inaccurate tracking of other similar color area farther away from the camera.

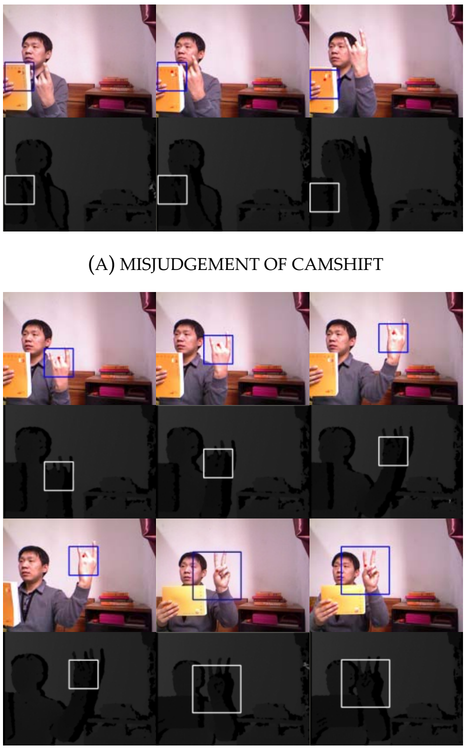
As shown in figure 2 (a), when using the CamShift, target may be lost sometimes. While (b) shows the tracking effect by DI_CamShift in the same situation (color images are frames from the Kinect color video, the corresponding depth image below are frames from the simultaneous depth video).



(A) CAMSHIFT LOST TRACKING GESTURE

(B) DI_CAMSHIFT ROBUST TRACKING

FIG. 2 COMPARE OF DI_CAMSHIFT AND CAMSHIFT



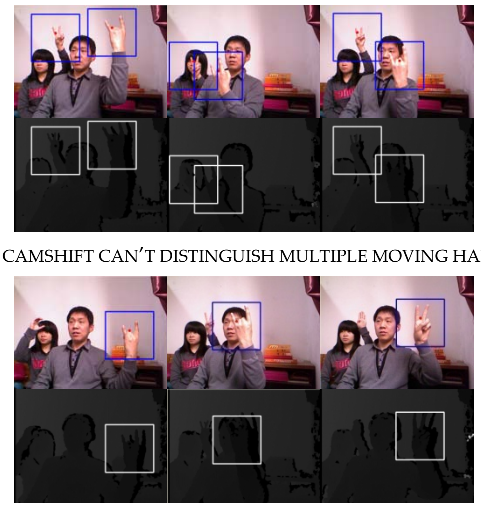
(A) MISJUDGEMENT OF CAMSHIFT

(B) ANTI-INTERFERENCE TRACKING OF DI_CAMSHIFT

FIG. 3 COMPARE OF DI_CAMSHIFT AND CAMSHIFT UNDER INTERFERENCE

In figure 3 (a), a book whose cover is yellow disturb the tracking because the color is similar to skin. When the hand closing to the yellow cover, Camshift has misjudging the cover to be the target hand and moving it's tracking window on the book. After the hand moving away from the book, it still stayed in place and keeping the false tracking. Affected by light and shooting, the book cover changed its color from canary yellow to bold yellow, but the DI_CamShift shows accurate tracking in (b) while the hand arriving closer and farther apart to the book.

At the time of tracking, sign gesture is considered to be the forefront part of human body, even there are more moving hands in video, only the closest to camera is the target for tracking. Camshift shows its tracking in figure 4 (a) when there are two moving gestures. Although located in the different distances and the posterior hand is an interference factor, the algorithm cannot distinguish and treat both of them as the tracking objects. (b) is a correct tracking by DI_Camshift which recognize the target by depth information.



(A) CAMSHIFT CAN'T DISTINGUISH MULTIPLE MOVING HANDS

(B) DI_CAMSHIFT CAN RECOGNIZE AND TRACKING THE REAL OBJECT

FIG. 4 COMPARE OF IDENTIFICATION ABILITY BETWEEN TWO ALGORITHMS

DI_Camshift presents the exact tracking under the circumstance of indoor and darker in figure 5. The last group of images in it shows the effective tracking in multi-hands video frame.

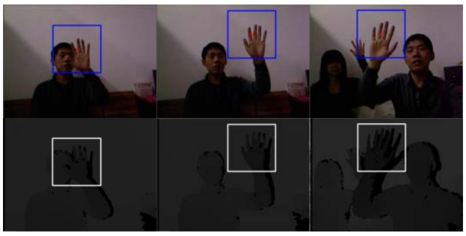


FIG. 5 TRACKING EFFECT OF DI_CAMSHIFT UNDER DARKER INDOOR CIRCUMSTANCE

SIFT-BoW Feature

SIFT (Scale Invariant Feature Transform) is a kind of operator used to describe local characteristics which extracts characteristic vectors unrelated to rotation, scale and location through searching extreme value point extraction in scale space. In the process of feature extraction, for each key point, 16 seed points (4 x 4) were used to describe it. Thus the obtained SIFT feature is a 128-dimensional vector while get all of the data from one key point in order to improve the robustness of sign language gesture matching, and SIFT vector got the scale and rotation invariance in this way.

Applied to sign language recognition, BoW (Bag of Words) characterize a sign language image through the way of regard it as a document collected several language visual vocabulary, and there is no ordering relation between different visual words. Different with text document, sign language visual word does not exist in an explicit style. Independent visual words of sign language must be extracted from the frame and generate BoW of it. This process mainly includes 4 steps as shown in figure 6.

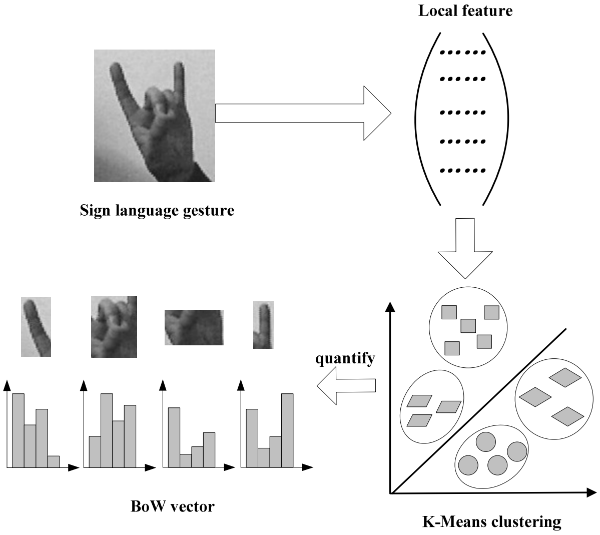


FIG. 6 4 STEPS OF BOW GENERATING

(1) Feature detection, getting interesting points by segmentation of sign language image.

(2) Feature presentation, using local feature descriptor of the image to represent its content.

(3) Generate visual words, generating SLVW (Sign Language Visual Word) from small region of the image described by local characteristics.

(4) Generate BoW, calculating frequency of the visual words in a sign language image and generate a

frequency histogram of visual word to represent the sign language gesture by BoW.

After calculating the SIFT feature, it was used as SLVW of gesture image and all of them were collected to build a sign language visual vocabulary based on K-Means clustering algorithm. According to the distance between visual vocabulary vectors, K-Means algorithm incorporates words of similar meaning to get K cluster centres as visual words to compose the basic vocabulary in word list, and then build the visual vocabulary by these visual words. Finally, words in the sign language visual vocabulary can be used to represent the corresponding image of sign language gesture. Refer to classification method of document, N nearest neighbour visual words were calculated for each vector in every sign language frames, and then measure the quality of the kth SLVW:

$$q(t_k) = \sum_{i=1}^{N}\sum_{j=1}^{M_i} \frac{1}{2^{i-1}} s\left(f_j, t_k\right) \tag{5}$$

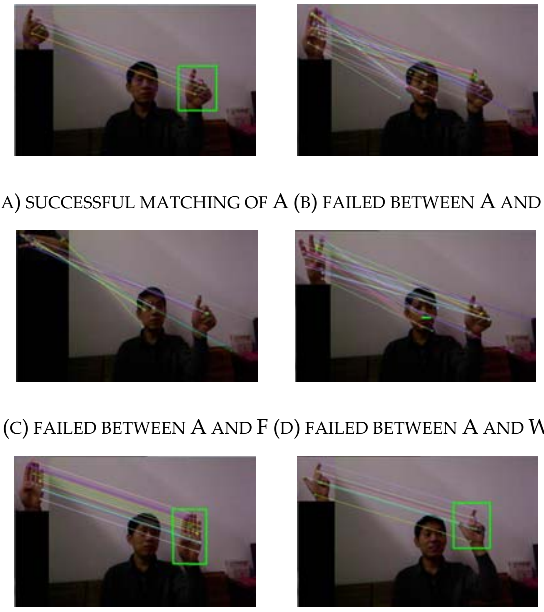
Where M_i is the number of eigenvector which is the ith close one to visual word t_k , $s(f_j,t_k)$ indicate the similarity between feature vector f_j and visual word t_k by Squared Euclidean Distance. For each visual word, it can be retained if the calculation of its quality satisfied the following formula, otherwise it will be abandoned.

$$q(t_k) =< h \max_{i=1,2,\ldots,K}(q(t_i)) \tag{6}$$

In (6), h is the quality threshold of visual word.

Experiments Based on SIFT-BoW Feature

All of the experiments captured sign language video by Kinect, and DI_CamShift was first used to hand tracking before calculate SIFT-BoW feature of each gesture in the search window, which is the key for gesture matching and recognition between prepared sign language repertoire and new coming gestures of sign language letter from the video. Using Squared Euclidean Distance, figure 7 shows the matching experiments of Chinese sign language letters based on its SIFT-BoW feature. (a) shows the successful matching between two images. The small gesture image at the top left is a static image of letter A random selected from the sign language repertoire, while the main picture is a frame of letter A in a sign language video from Kinect. Each of (b) (c) (d) shows failed matching when the gesture is letter A in video and using SIFT-BoW feature of static letter B, F, W to match it. (e) gives the correct matching of letter B and (f) recognizes the letter L. Comparing the experimental results, SIFT-BoW feature is an effective and robust feature for sign language recognition research.



(A) SUCCESSFUL MATCHING OF A (B) FAILED BETWEEN A AND B

(C) FAILED BETWEEN A AND F (D) FAILED BETWEEN A AND W

(E) SUCCESSFUL MATCHING OF B (F) SUCCESSFUL MATCHING OF L

FIG. 7 SIGN LANGUAGE MATCHING BASED ON SIFT-BOW

Conclusions

Both of an improved CamShift algorithm based on depth image information from Kinect sign language video and SIFT-BoW feature were used for Chinese sign language recognition research in this paper. The DI_CamShift algorithm adjusts the size of Search Window by calculating the spindle direction and mass center of sign gestures in depth images. It can make the steady gesture tracking continuously and improved disadvantages of CamShift in lost tracking target and misjudgement with similar color. Based on accurate tracking of hand, SIFT-BoW feature is calculated from the search window for matching and recognition. Comparative experiments show that the above method has a strong adaptive ability for varieties of scene, be able to adjust search scope timely, improved speed of tracking and achieved high efficiency and robustness gesture matching.

ACKNOWLEDGMENT

This work is supported by Xi’an University of Arts and Science, young and Middle-aged professional scientific research project (No. 90147).

REFERENCES

Juan, Pablo Wachs, Mathias, Kolsch, Helman, Stern, et al. Vision-Based Hand-Gesture Applications [J]. Communications of the ACM, February 2011, Vol. 54 No.2.

LIU, Yang-wen, HUO, Hong, FANG, Tao. Visual Words Ambiguity Analysis in BOW Model [J]. Computer Engineering. Vol37, No. 19, 2011. 10.

Morel, J M, Yu, G S. ASIFT: A New Framework for Fully Affine Invariant Image Comparison [J]. Society for Industrial and Applied Mathematics Journal on Image Sciences, 2009, 2(2): 438-469.

OpenKinect Organization. Imaging Information for Kinect [EB/OL]. http://openkinect.org/wiki/Imaging_Information, 2013.

RAHEJA, J.L, CHAUDHARV, A, SIGNAL, K. Tracking of Fingertips and Centres of Palm using Kinect [A]. CIMSiM [C], Langkawi, IEEE, 2011:248-252.

WANG, Yu-shi, GAO, Wen. Kernel-based Image Classification Using the Context of Visual Words [J]. Journal of Image and Graphics, 2010.4.

ZHANG, Qiu-yu, WANG, Dao-dong, ZHANG, Mo-yi, et al. Hand Gesture Recognition Based on Bag of Features and Support Vector Machine [J]. Journal of Computer Applications, 2012, 32(12):3329-3396.